Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal. Vol. 2, Issue 1 (2009), 37 – 57
ISSN 2070-5948, DOI 10.1285/i20705948v2n1p37

CHRUN RISK MITIGATION MODELS FOR STUDENT'S BEHAVIOR

Silvia Figini*, Emanuele De Quarti, Paolo Giudici

Department of Statistics and Applied Economics L. Lenti, University of Pavia, Pavia, Italy

Received 22 June 2009; Accepted 10 August 2009
Available online 15 August 2009

Abstract: *In order to prevent the phenomenon of abandonment (churn), the objective of this paper is to analyze university student's careers. The results of our research will be used to plan activities of collective or single tutoring and to measure the efficiency and efficacy for specific courses. In particular, the analysis considers years of history for two faculties in the University of Pavia: psychology and biology. In order to estimate for each student a measure of churn risk, our methodological approach is based on a duration predictive model. Empirical evidences are given on the basis of a real data set.*

Keywords. *Predictive duration model, Cox Model, Hazard risk, Churn*.

1 Introduction

The objective of this paper is to analyze university student's careers and to evaluate for each student the churn risk. The analysis considers two faculties in the University of Pavia: psychology and biology. Psychology has been chosen because of the limited access to the enrolment. For this reason we believe to deal with strongly motivated students; this fact makes the analysis of the data particularly meaningful. On the other hand, biology is subject to a particular phenomenon: some students take exams in this faculty, but they hope to pass to the faculty of medicine the following year.
This research proposes statistical duration models useful for churn evaluation, in order to reduce the number of students who leave their university career without reaching any degree.
In the business field is rather immediate to establish the number of clients that abandons a service. The distinction is usually made between voluntary and involuntary churn. Involuntary churn occurs when the company terminates the customers' contract or account - usually on the basis of a poor payment history. Voluntary churn is when the customer decides to take their business elsewhere.

* Corresponding Author. Email: silvia.figini@unipv.it

Now we adapt these concepts to a situation concerning university students. Voluntary churn happens when the student interrupts his studies in a University. The student may definitely leave his/her career: we will call this possibility "renounced student". On the other hand, the student may decide to start again his career in another University: we will call this possibility "dismissed student". "Renounced student" and "dismissed student" are positions which are officially enacted by documents, compiled in the central reception office of the university.
There are also positions of students that officially result "active", but that don't take exams anymore and they give up paying the annual tax for years. The reception office considers these kind of students as "renounced student" after eight years of complete inactivity, but it is clearly possible to establish in advance this condition of abandonment.
The problem of student education and evaluation is well recognised, and relevant contributions in this area of research are given by Dutton et al., 2005, Dutton et al., 2001, Simonoff, 1997, Spooner et al., 1999, Wallace et al., 1997 and more recently by Carnell, 2008.
Considering the problem at hand, it is necessary to individualize a criterion to discriminate between "active" and "inactive" students. The missed payment of the annual tax of registration, except for graduated students, highlights clearly a possible risk of churn. It is necessary then to verify that the student does not restart to pay the tax in following years.
As results, the statistical approach employed defines profiles of students with a high churn risk, evaluating the dependence of risks on the basis of the following factors: credits, average mark of exams and social-demographic variables such as gender, date of birth, province of residence, and type of middle school diploma. The resulting profiles could be used for beginning tutoring activities in order to prevent churn risk.
This paper is organised as follows: Section 2 presents the available data, Section 3 describes our methodological proposal for churn risk estimation; finally Section 4 reports the empirical evidence on the basis of a real data set. Section 5 summarises the research and highlights the conclusions and further ideas of research.

2. Data Set

The available data for our analysis contain information about the faculties of Psychology and Biology. The samples at hand for Psychology and Biology are composed of the total amount of students in the two faculties. The statistical unit of interest in the two data sets is the student.
The number of students under consideration is 845 in the faculty of Psychology and 1037 in the faculty of Biology. The period of time considered is: 2001-2007.

The independent variables collected for both faculties are:

- *Id*: student identification number
- *Date of birth*: year of birth
- *Gender*: female or male
- *Prov Rec*: Italian province of residence
- *Diploma*: type of middle school diploma (professional institute, technical institute, classical high school, linguistic high school, scientific high school, teacher's college, other).
- *Position*: current position of the student (Active, Dismissed, Interruption, Graduated, Renounced, Declined).
- *Credits*: credits maturated by the student in the exams for each year of his career. The

minimum value is 5, the maximum is 185.

- *Average Mark*: average mark of the given exams for each year of the career. The minimum value is 18, the maximum is 30.
- *Tax*: type of payment of the tax for each year (none, only first, full).

Concerning the quantitative variables, Table 1 reports descriptive statistics for Psychology and Biology, such as the range, the standard deviation, the mean and the coefficient of variation.

Table 1: Descriptive statistics for the quantitative variables

Descriptive Measure	Credits Psychology	Credits Biology	Average Mark Psychology	Average Mark Biology
Range	180	171	12	12
Standard deviation	38,78	62,03	2,24	2,54
Mean	114,37	72,36	25,12	24,65
Coefficient of variation	33,9%	85,7%	8,9%	10,3%

Missing values are present for the "Credits" and the "Average mark" variables and it means that the student didn't take any exam in that year under consideration.
The target variable is called "Churn". The status variable "Churn" is binary and it presents these values:

- 1: event has occurred;
- 0: censored cases.

Table 2 reports the distribution frequency for the target variable in the faculties under analysis.

Table 2: Frequency distribution for the target variable

Churn	Psychology	Biology
Events (Churn=1)	188	349
Non Events (Churn=0)	657	688
Total	845	1037

We will use it to examine the distribution of times between two events, the beginning and the end of the career. This kind of data includes some censored cases. Censored cases are cases for which the second event is not recorded (Churn=0). In this case the censored cases include graduated students and "active" students, who keep on with their university career and paying the tax.

3. Methodological Proposal

Before illustrating our methodological proposal, in our opinion concerning the causes of churn for the student, it is possible to identify a number of components that can generate such behaviour:

- a static component, determined by the characteristics of the students;
- a dynamic component, that encloses trend and the contacts of the students with the university career;

- a seasonal part, linked to the period of exams;
- external factors.

The goal of the university is to identify students that are likely to leave and join a new university or a job. This objective is well perceived by the university top management, which considers lowering churn one of the key targets.
The churn models currently used to predict churn is a logistic regression compared with a classification tree. Tree models can be defined as a recursive procedure, through which a set of "n" statistical units is progressively divided in groups, according to a divisive rule which aims to maximize a homogeneity or purity measure of the response variable in each of the obtained groups. At each step of the procedure, a divisive rule is specified by: the choice of an explanatory variable to split, the choice of a splitting rule for such variable, which establishes how to part the observations.
The main result of a tree model is a final partition of the observations: to achieve this it is necessary to specify stopping criteria for the divisive process.
Tree models may show problems in time-dependent applications, such as churn applications. The same holds for logistic regression.
In order to obtain a predictive tool which is able to consider the fact that churn data is ordered in calendar time, the use of new methods is necessary. We can sum up at least three main weak points of traditional models in our set-up, which are all related to time-dependence:

- redundancy of information: the database contains variables which gives redundant information; as these variables are time dependent, they may induce biased effects in the final estimates;
- presence of fragmentary information, depending on the measurement time;
- excessive weight of the different temporal perspectives (the method used to build predictive models cannot catch this temporal dimension).

The previous points explain why it is required to search a novel methodology to predict churn. This research proposes to use the survival analysis approach: this method was born in the medical field, but in the university we applied it in a new way to predict churn behavior.

3.1 Survival analysis models to estimate churn

We now turn our attention towards the application of methodologies aimed at modeling survival risks. In our case the study of risk concerns the value that derives from the loss of a student. The goal is to determine which combination of covariates affects the risk function, studying specifically the characteristics and the relation with the probability of survival for every student.
Survival analysis is concerned with studying the time between starting a study and a subsequent event (churn). All of the standard approaches to survival analysis are probabilistic or stochastic. That is, the times at which events occur are assumed to be realizations of some random processes. It follows that T, the event time for some particular individual, is a random variable having a probability distribution.
A useful, model-free approach for all random variables is nonparametric, that is, using the cumulative distribution function. The cumulative distribution function of a variable T, denoted by F(t), is a function that tells us the probability that the variable will be less than or equal to any

value t that we choose. Thus, $F(t) = P\{T \leq t\}$. If we know the value of F for every value of t, then we know everything about the distribution of T. In survival analysis it is more common to work with a closely related function called the survivor function defined as $S(t) = P\{T>t\} = 1 - F(t)$. If the event of interest is a death (or, equivalently, a churn) the survivor function gives the probability of surviving beyond t. Because S is a probability we know that it is bounded by 0 and 1 and because T cannot be negative, we know that S(0) =1. Finally, as t gets larger, S never increases. Often the objective is to compare survival functions for different subgroups in a sample (according to different factors: Gender, Average mark, Date of birth, Italian province of residence, Diploma). If the survival function for one group is always higher than the survival function for another group, then the first group clearly lives longer than the second group.
When variables are continuous, another common way of describing their probability distributions is the probability density function. This function is defined as:

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$$

Or better, the probability density function is just the derivative or slope of the cumulative distribution function. For continuous survival data, the hazard function is actually more popular than the probability density function as a way of describing distributions. The hazard function is defined as:

$$h(t) = \lim_{\varepsilon t \to 0} \frac{\Pr\ t \leq T \leq t + \varepsilon t \mid T \geq t}{\varepsilon t}$$

The aim of the definition is to quantify the instantaneous risk that an event will occur at time t. Since time is continuous, the probability that an event will occur exactly at that time t is necessarily 0. But we can talk about the probability that an event occurs in the small interval between t and t+ εt and we also want to make this probability conditional on the student surviving to time t. For this formulation the hazard function is sometimes described as a conditional density and, when events are repeatable, the hazard function is often referred to as the intensity function. The survival function, the probability density function and the hazard function are equivalent ways of describing a continuous probability distribution. Another formula expresses the hazard in terms of the probability density function:

$$h(t) = \frac{f(t)}{S(t)}$$

and together previous equations imply that

$$h(t) = -\frac{d}{dt} \log S(t)$$

Integrating both sides of equation above gives an expression for the survival function in terms of the hazard function:

$$S(t) = \exp\left(-\int_0^t h(u)\,du\right)$$

With regard to numerical magnitude, the hazard is a dimensional quantity that has the form: number of events per interval of time. The interpretation of the hazard as the expected numbers of events in a one-unit interval of time makes sense when events are repeatable. The available database for our analysis contains information that can affect the distribution of the event time, as the demographic variables, or variables about the career, the payment, the contacts, geographical origin and curricula.
In this application, the target variable has a temporal nature and, for this reason, it is preferable to build predictive models through survival analysis.
The actual advantages of using a survival analysis approach compared with a traditional one can be summarised as follows:

- to correctly align the students regarding their cycle of life in the university;
- to analyze the real behaviour of the students churn.

In order to build a survival analysis model, two variables are required: one variable of status (distinguish between active and non active students) and one of duration (indicator of student seniority) . The first step in the analysis of survival data (for the descriptive study) consists in a plot of the survival function and the risk. The survival function is estimated through the methodology of Kaplan Meier (see e.g. Kaplan and Meier, 1958). The Kaplan Meier estimator is the most widely used method for estimating a survival function and it is based on a nonparametric maximum likelihood estimator. When there are non censored data the KM estimator is just the sample proportion of observations with event times greater than t.
The situation is also quite simple in the case of single right censoring, that is, when all the censored cases are censored at the same time c and all the observed event time are less than c. In that case, for all $t \leq c$ the KM estimator is still the sample proportion of observations with event times longer than t. Things get more complicated when some censoring times are smaller then some event times. In that instance, the observed proportion of cases with event times longer than t can be biased downward because cases that are censored before t may, in fact, be "dead" before t without our knowledge. The solution is as follows. Suppose there are K distinct event times, $t_1 < t_2 < \ldots < t_k$. At each time t_j there are n_j individuals who are told to be at risk of an event. To be at risk means they have not experienced an event not being censored before that time t_j . If any cases are censored at exactly t_j , there are also considered to be at risk at t_j . Let d_j be the number of individuals who die at time t_j . The KM estimator is defined as:

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left[1 - \frac{d_j}{n_j}\right], \text{ for } t_1 \leq t \leq t_k.$$

This formula says that, for a given time t, take all the event times that are less than or equal to t. For each of those event times, compute the quantity in brackets, which can be interpreted as the conditional probability of surviving to time t_{j+1} , given that one has survived to time t_j . Then multiply all of these survival probabilities together.

In order to build a predictive model a natural choice is to implement Cox's model (see e.g. Cox, 1972). Cox made two significant innovations. First he proposed a model that is a proportional hazards model. Second he proposed a new estimation method that was later named partial likelihood or more accurately, maximum partial likelihood. Our analysis start with the basic model that does not include time-dependent covariate or non proportional hazards. The model is usually written as:

The previous equation says that the hazard for individual i at time t is the product of two factors: a baseline hazard function that is left unspecified, and a linear combination of a set of p fixed covariates, which is then exponentiated. The baseline function can be regarded as the hazard function for an individual whose covariates all have values 0. The model is called proportional hazard model because the hazard for any individual is a fixed proportion of the hazard for any other individual. To see this, take the ratio of the hazards for two individuals I and j:

$$\frac{h_i(t)}{h_j(t)} = \exp[\beta_1(x_{i1} - x_{j1}) + ... + \beta_p(x_{ip} - x_{jp})]$$

What is important about this equation is that the baseline cancels out of the numerator and denominator. As a result, the ratio of the hazards is constant over time.
In Cox model building the objective is to identify the variables that are more associated with the churn event. This implies that a model selection exercise, aimed at choosing the statistical model that best fits the data, is to be carried out. The statistical literature presents many references for model selection (see e.g. Giudici, 2003).

A very important remark is that Cox model generates survival functions that are adjusted for covariate values. More precisely, the survival function is computed according to the following:

Once a Cox model has been fitted, it is advisable to produce diagnostic statistics, based on the analysis of residuals, to verify if the hypotheses underlying the model are correct. In our case they were found to be correct, so we could proceed with the predictive stage. In the prediction step the effectiveness of the model will be evaluated in terms of predictive accuracy following cross validation.

4. Empirical evidence

4.1 Exploratory and descriptive results

Through the analysis of the payments of the taxes and the activities of the students we classify the careers, defining the variable churn, that points out abandonment. The frequency distribution for the variable churn in the faculties of Psychology and Biology, is reported in Table 2. From

Table 2, in Psychology 188 students experiment the event abandonment, while 657 students, 77,8% of the observations, are censored cases.
On the other hand in Biology 349 students experiment the event abandonment, while 688 students, 66,3% of the observations, are censored cases.
In order to estimate the survival function the empirical analysis uses the Kaplan Maier estimator described in Section 3. The results achieved in Table 3, are based only two variables: the churn variable and the time (yearly duration time).

Table 3: Kaplan Meier estimation

time	n.risk (Psychology)	n.event (Psychology)	Survival (Psychology)	n.risk (Biology)	n.event (Biology)	Survival (Biology)
1	845	37	0.956	1037	303	0.708
2	707	44	0.897	517	30	0.667
3	448	57	0.783	368	9	0.65
4	96	37	0.481	116	6	0.617
5	21	13	0.183	46	1	0.603

Table 3 shows the results in terms of survival probabilities estimated for the faculty of psychology and biology. In particular, for each duration time, Table 3 reports the number of cases at risk, the number of events and the relative survival probability.

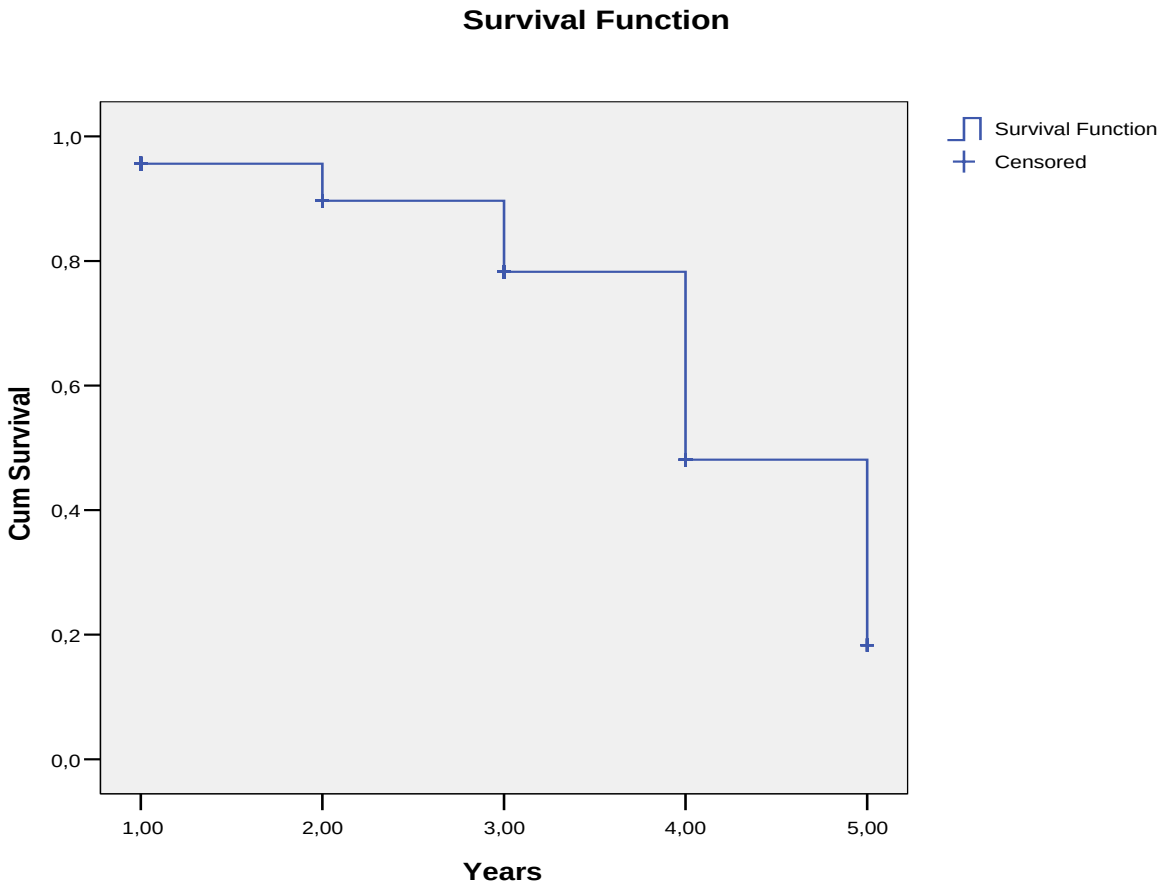

Figure 1: Survival probability function - Psychology

As we can observe from Table 3, the total number of students in Psychology under observations is equal to 845 with a probability of survival equal to 0.956. 37 students experiment in the first year the event of interest. After 5 years the survival probability is equal to 0.183 and the number of students at risk is equal to 21. Focusing on Biology, the strong rate of abandonment of the first year engraves the probabilities of survival, equal after one year to 0.708 for a total of 303 students that experiment the event.
Nevertheless, in the following years the situation has the tendency to settle since the probabilities of survival fluctuate between 0.667 of the second year and 0.603 of the fifth year.
Considering the faculty of Psychology the results in Table 3 are summarised in Figure 1. In Figure 1 the x-axis indicate lifetime and the y-axis the survival probabilities.
From Figure 1 it is interesting to notice that the survival function has varying slopes, corresponding to different periods. When the curve decreases rapidly we have time periods with high churn rates; when the curve decreases softly we have periods of "loyalty". We remark that the final jump is due to a distortion caused by a few data, in the tail of the lifecycle distribution.
Starting from Figure 1, is possible to derive the hazard risk curve, reported in Figure 2. In Figure 2 the x-axis indicate lifetime and the y-axis the hazard risk.

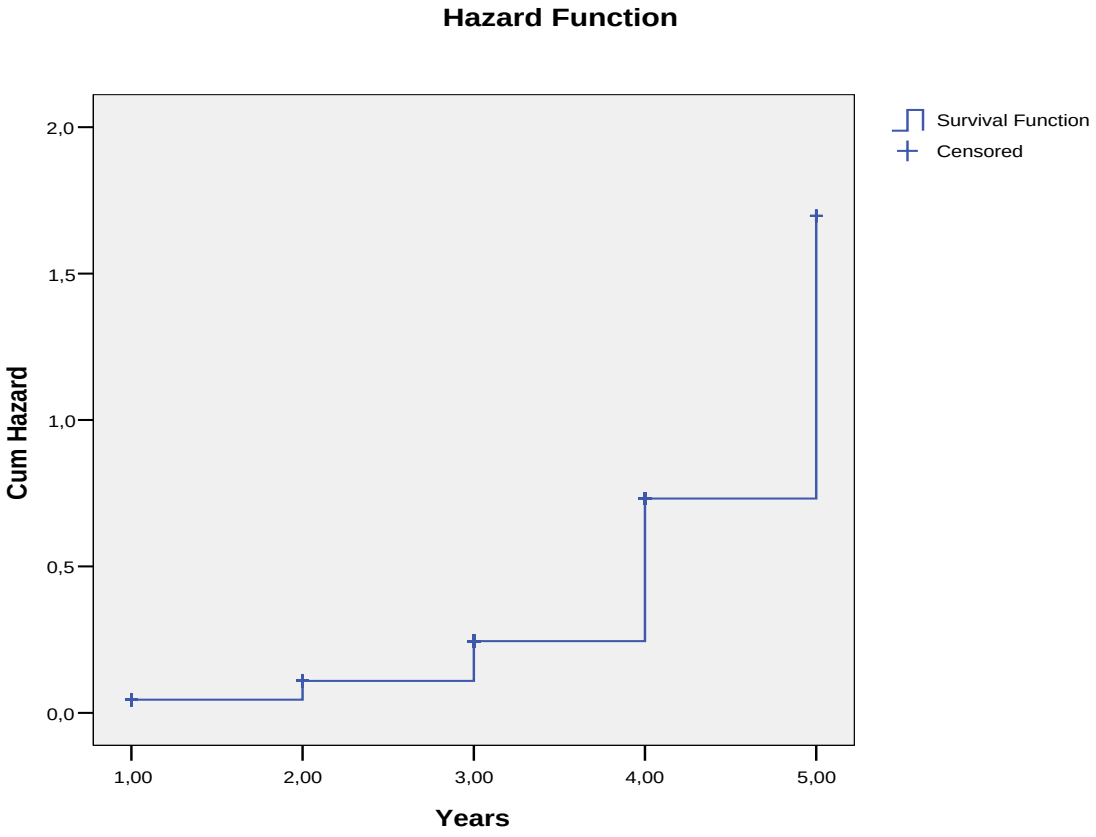

Figure 2: Hazard function - Psychology

The hazard risk increases moderately in the second and third year. Students subject to risk increase in more evident way starting from the fourth year. A very useful information, is the

calculation of the life expectancy of the students. This can be obtained as a sum over all observed event times:

$$\hat{S}(t_{(j)})(t_{(j)} - t_{(j-1)}),$$

where $\hat{S}(t_{(j)})$ is the estimate of the survival function at the j-th event time, obtained using Kaplan Meier method, and *t* is a duration indicator. ($t_{(0)}$ is by assumption equal to 0). We remark that life expectancy tends to be underestimated if most observed event types are censored (i.e., no more observable).
In Figure 3, we plot the survival function, of the students of biology in the six analyzed years.

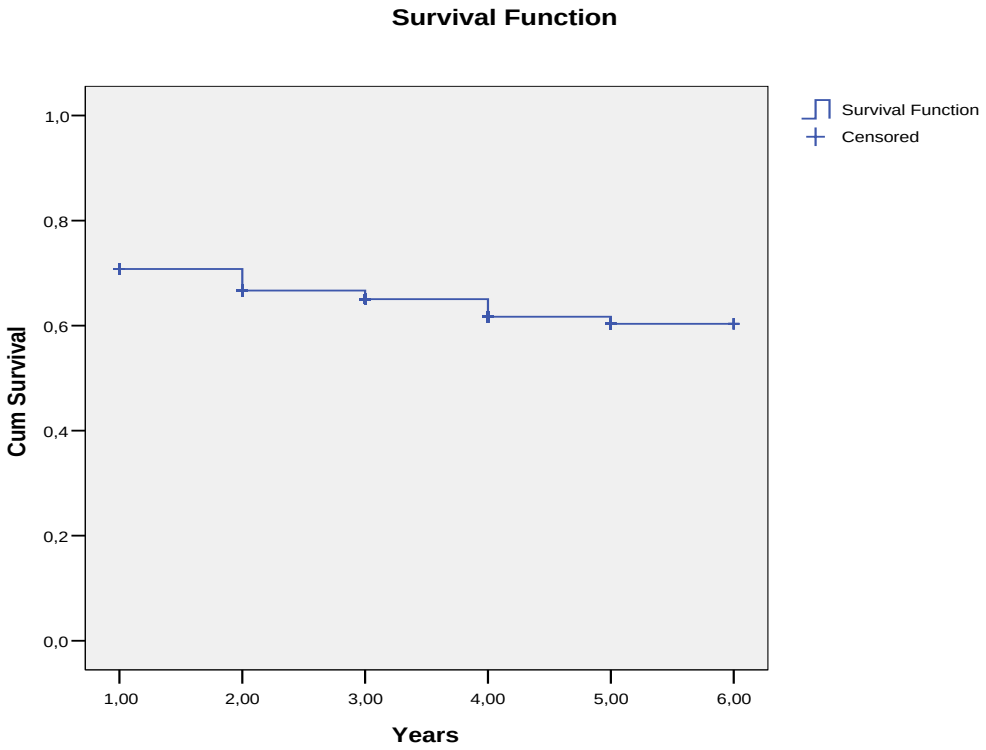

Figure 3: Survival probability function - Biology

The curve decreases rapidly only after one year, so we have an high churn rate. In the other duration time the curve decreases softly: they are periods of "loyalty".
The hazard risk curve is reported in Figure 4. The hazard risk moderately increases after the first and the third year. As we can observe from Figure 4, the hazard risk shows high levels in the first year and confirms empirical evidence shown in Figure 3.

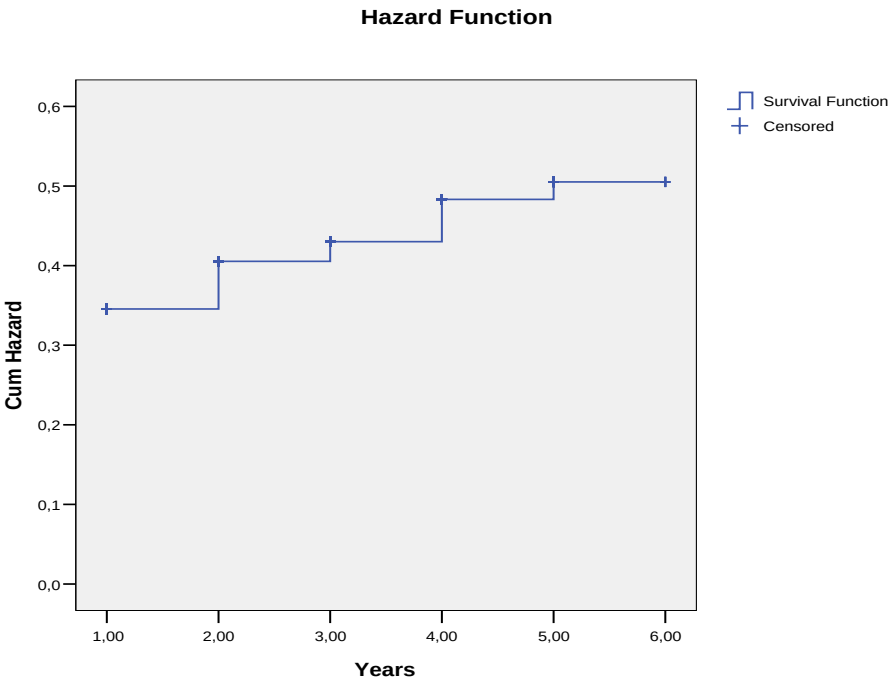

Figure 4: Hazard function - Biology

We now move to the descriptive comparison of hazard and/or survival curves, according to the available covariates described in Section 2: Gender, Average mark, Date of birth, Italian province of residence and Diploma.

On the basis of the available sample, the performance of female students is better than the male colleagues, as shown for both faculties in Table 4. In the analysis we consider 143 males and 702 females.

In particular, from Table 4 the survival probability for the females after 1 year duration time is equal to 0.966 and 0.909 for the males. The females show lowest churn risk. In fact the faculty of psychology in our university is preferred by female's students. This thesis is very realistic considering also the distribution frequency for the variable gender (the sample size of psychology is composed of 143 males and 702 females). Figure 5 shows an example of such comparison, on the basis of the gender, for the Faculty of Psychology.

Table 4: Kaplan Meier estimation standing to Gender

Gender=Male						
time	**n.risk (Psychology)**	**n.event (Psychology)**	**Survival (Psychology)**	**n.risk (Biology)**	**n.event (Biology)**	**Survival (Biology)**
1	143	13	0.909	316	95	0.699
2	113	15	0.788	153	15	0.631
3	73	15	0.626	101	1	0.625
4	16	7	0.352	42	3	0.58
5	4	4	0	17	1	0.546
Gender=Female						
time	**n.risk (Psychology)**	**n.event (Psychology)**	**Survival (Psychology)**	**n.risk (Biology)**	**n.event (Biology)**	**Survival (Biology)**
1	702	24	0.966	721	208	0.712
2	594	29	0.919	364	15	0.682
3	375	42	0.816	267	8	0.662
4	80	30	0.51	74	3	0.635
5	17	9	0.24	721	208	0.712

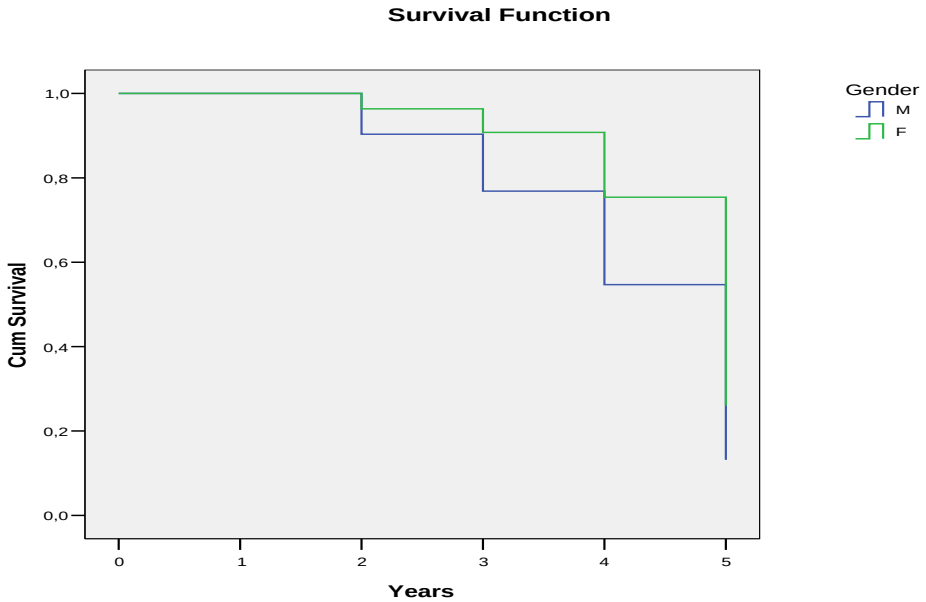

Figure 5: Survival function, standing to Gender - Psychology

Figure 6 shows the same comparison, for the faculty of Biology.

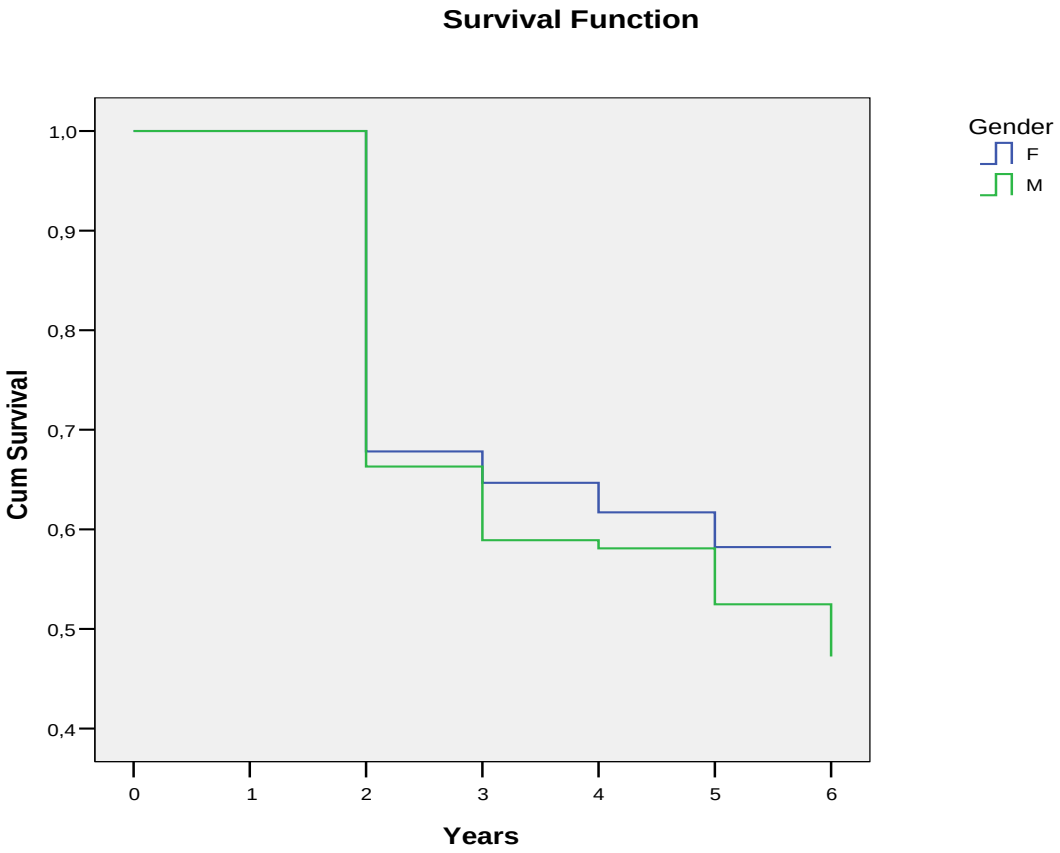

Figure 6: Survival function, standing to Gender - Biology

The survival probabilities estimated are quite similar for males and females in the first years duration time. As for the faculty of psychology, the performance of the female students is better than the male colleagues. We have analyzed therefore the performances of students coming from different Italian provinces. Examining the data set of psychology, we notice that the greater part of the students comes from the same region where the faculty is situated, Lombardy. The students resident in the provinces of "Alessandria" and "Sondrio" have underlined very positive results, while the ones from the province of "Genoa" have negative results.

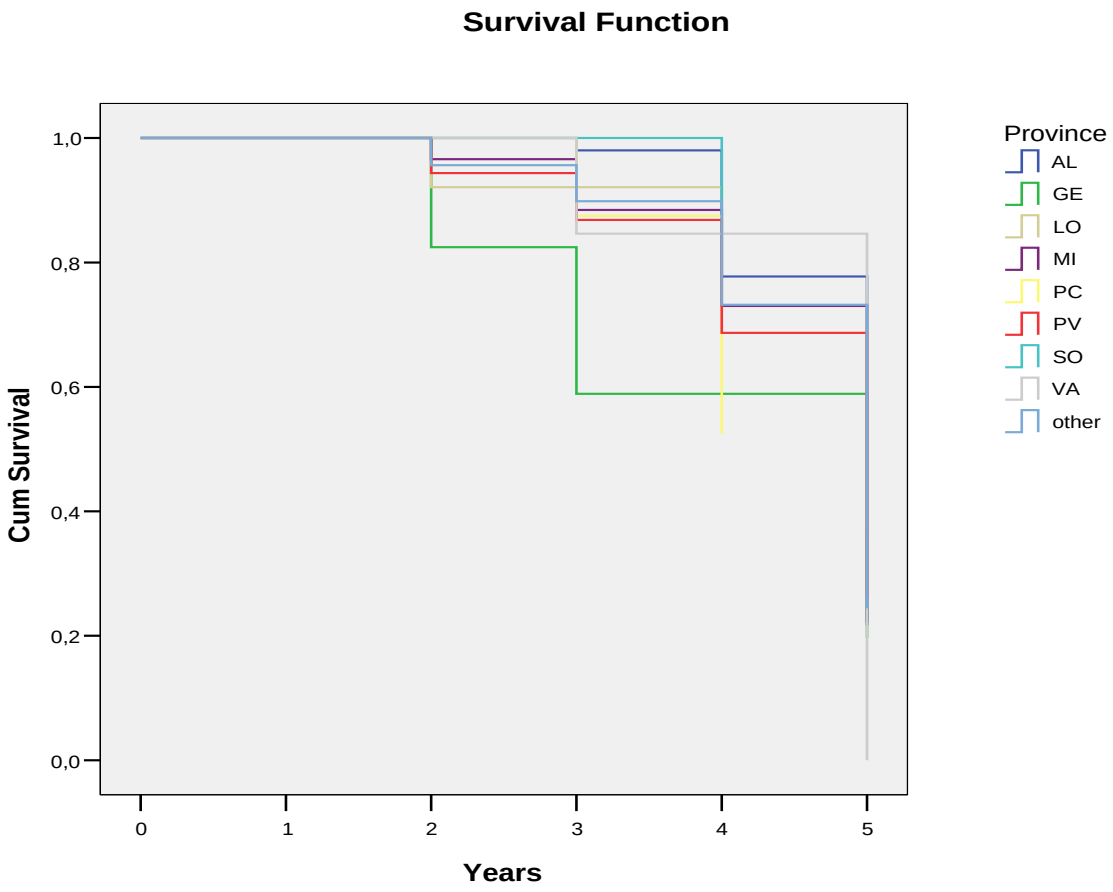

Figure 7: Survival function, standing to Province - Psychology

From figure 7 we can appreciate significant differences in survival curves, due to the geographical factor. These differences have been further confirmed by a formal test of hypotheses (Scheffè multiple comparison test, see e.g. Singer and Willet, 2003). Conclusions similar to those in Figure 7 can be obtained for all explanatory variables; in our experience this represents a great wealth for university usages.

Figure 8 shows the results for the faculty of Biology: the analysis of the groups coming from different provinces shows that the courses of the careers are rather similar. Students coming from "Agrigento", "Lodi" and "Pavia" present better careers after one year, with probabilities respectively of the 0.792, 0.769, 0.727.

We derive now the survival probability distribution function (Figure 9) considering the diploma of high school as strata variable.

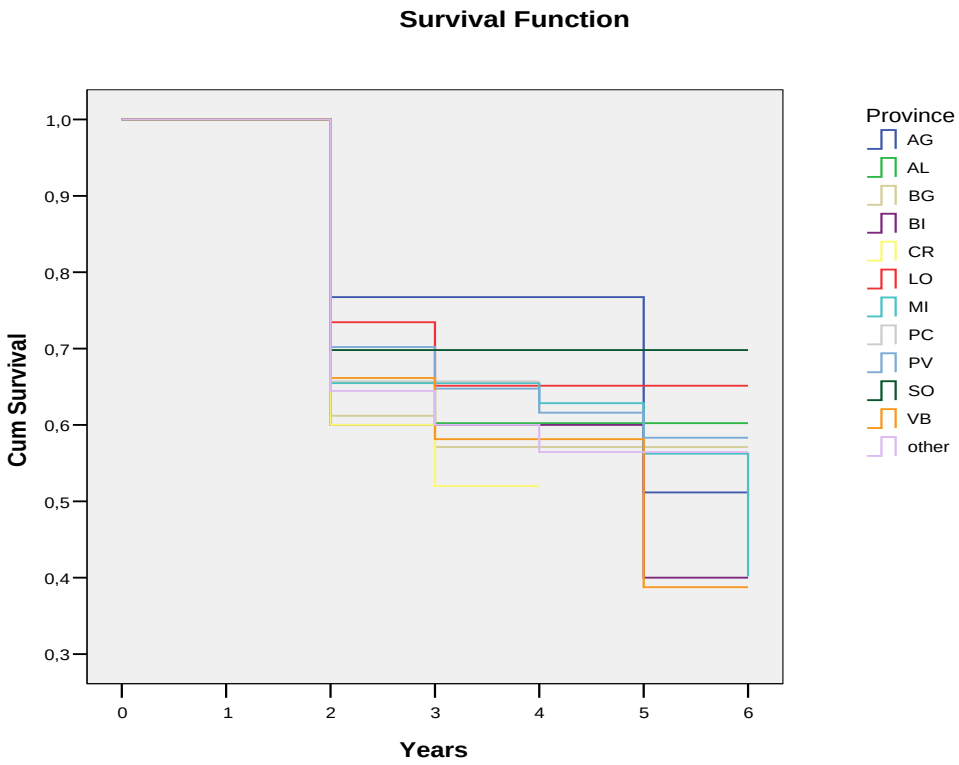

Figure 8: Survival function, standing to Province - Biology

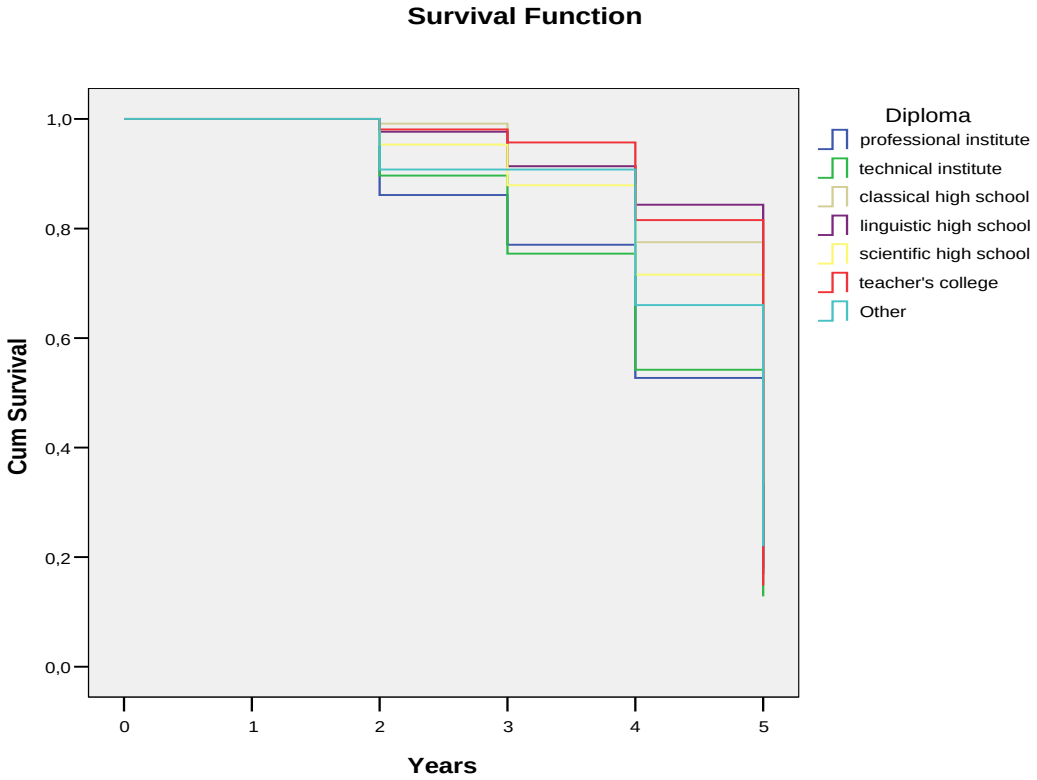

Figure 9: Survival function, standing to Diploma - Psychology

As it regards the diploma of senior high school, in the faculty of Psychology, the best results have been noticed for the students coming from classical (128 students), linguistic (47 students) and scientific (340 students) high school and teacher's college. The survival probabilities after one year duration time are respectively: 0.992 for classical school, 0.976 for linguistic school and 0.954 for the scientific school.

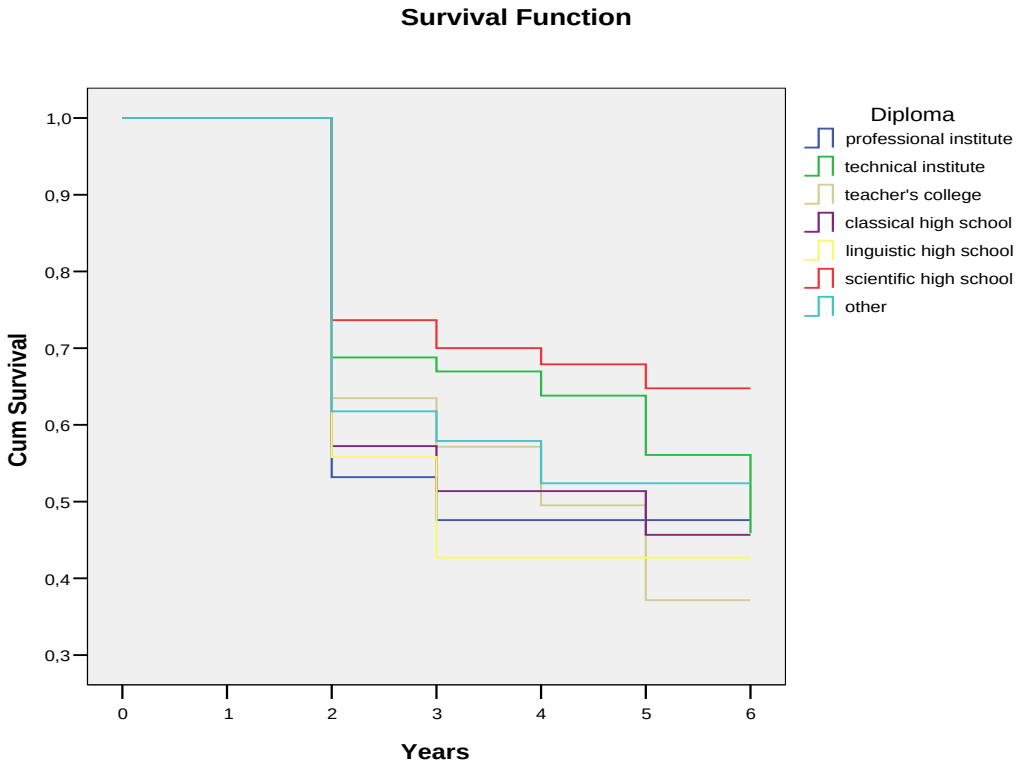

Figure 10: Survival function, standing to Diploma - Biology

In the Faculty of biology, students coming from scientific high school and technical institutes present appreciable results. Particularly the probabilities of survival for the students coming from a scientific high school (517 students) is closed to 0.7 for the considered first 4 years. For the technical institute (152 students) the survival probability starts at 0.711 and after 4 years duration time is equal to 0.60. Very difficult situations in terms of churn risk are present for the students coming from classical high school (163 students). In fact, the faculty of biology required, as background, a good knowledge for mathematics, science and topics related to medicine. Those issues are popular in scientific high schools and in specific technical institutes.

4.2 Inferential results

We now move to the building of a full predictive model based on the Cox's model described in Section 3. Cox made two significant innovations. First he proposed a model that is a proportional hazard model. Second he proposed a new estimation method that was later named partial likelihood or more accurately, maximum partial likelihood.

The model is called proportional hazard model because the hazard for any individual is a fixed proportion of the hazard for any other individual.

In Cox's model building the objective is to identify the variables that are more associated with the churn event. This implies that a model selection exercise, aimed at choosing the statistical model that best fits the data, is to be carried out.
We will start with the basic model that does not include time-dependent covariate or non proportional hazards. As covariates we consider: gender, province of residence, diploma, average mark, sustained credits.
The result of the procedure is a set of five explanatory variables. Such variables can be grouped in three main categories, according to the sign of their association with the churn rate, represented by the hazard ratio.

Table 5: Cox's Model analysis - Psychology

Variable	Estimates	Standard error	z	p-value	Hazard ratio	lower 0.95	upper 0.95
Gender	-0.3698	0.169	-2.188	0.029	0.691	0.496	0.962
prov_rec	0.0358	0.0326	1.099	0.27	1.036	0.972	1.105
Diploma	-0.1093	0.0417	-2.622	0.0087	0.896	0.826	0.973
average_mark	0.0404	0.0743	0.543	0.59	1.041	0.9	1.204
Credits_banded	-1.1024	0.0627	-17.591	0.0001	0.332	0.294	0.375

In Table 5 we report the results for the Cox model in the faculty of psychology. The results show significant variables as diploma and credits. In more details, we group the variables taking into account the association with the event of interest. In particular, on the basis of the hazard ration in Table 5, we observe:

- variables that show a negative association with churn (e.g. Gender, Diploma, Credits_banded);
- variables that have no association with churn (e.g. prov_rec, average_mark).

We remark that the analysis reported in this section and specifically starting form Table 5 are more precisely respect to the evidence reported in Section 4.1 because are based on a multivariate statistical analysis.
Considering the faculty of biology, the results for the Cox model are reported in Table 6. The significant variable is credits.
Based on Table 6, the faculty of Biology presents:

- variables that show a negative association (e.g. Gender, prov_rec, Credits banded, Diploma)
- variables that have no association (e.g., average mark).

Table 6: Cox's Model analysis - Biology

Variable	Estimates	Standard error	z	p-value	Hazard ratio	lower 0.95	upper 0.95
Gender	-0.0591	0.1148	-0.515	0.61	0.943	0.753	1.18
prov_rec	-0.0046	0.0175	-0.262	0.79	0.995	0.962	1.03
Diploma	-0.0348	0.0303	-1.148	0.25	0.966	0.91	1.02
average_mark	0.06433	0.0399	1.612	0.11	1.066	0.986	1.15
Credits_banded	-1.1899	0.0711	-16.73	0.0001	0.304	0.265	0.35

Once a Cox model has been fitted, it is advisable to produce diagnostic statistics, based on the analysis of residuals, to verify if the hypotheses underlying the model are correct. In our case they were found to be correct, so we could proceed with the predictive stage.

4.3 Predictive Model Assessment

In the prediction step the performance of the model will be evaluated in terms of predictive accuracy.

The analysis splits the datasets in the two usual subsets: training (60% of the observations) and test (40% of the observations). Both samples have been proportionally sampled, with respect to the target variable.

We have focused our attention on the 2 years ahead prediction, in order to evaluate the predictive performance of the model, and compare it with classical data mining models.

Once survival probabilities have been calculated, we have implemented a procedure to build the confusion matrix and, correspondingly, the percentage of captured true churners of the model. Table 5 contains the results of such comparison; in correspondence of each estimated probability decile, we report the percentage true churners in it (% Captured Response Rate).

Considering the data at hand and the frequency in each percentile, the model captured very well the event of interest as shown in Table 7.

Table 7: Captured Response

Percentile	Captured Response Rate (Psychology)	Captured Response Rate (Biology)	Cumulative Captured Response Rate (Psychology)	Cumulative Captured Response Rate (Biology)
1	12.86%	15.48%	12.86%	15.48%
2	12.86%	15.48%	25.72%	30.96%
3	12.36%	14.73%	38.08%	45.69%
4	11.85%	13.61%	49.93%	59.3%
5	11.47%	13.43%	61.4%	72.73%
6	10.71%	5.97%	72.11%	78.7%
7	9.56%	8.39%	81.67%	87.09%
8	9.01%	4.66%	90.68%	91.75%
9	6.8%	3.73%	97.48%	95.48%
10	2.52%	4.52%	100%	100%

From Table 7 note that, the first deciles contains the student with the highest estimated churn probability and the percentage lowers down in subsequent deciles, thus giving an overall picture of the good performance of the model.
However, we remark that, differently from what occurred with classical models, the students with the highest estimated churn rate are now not necessarily those whose contact is close to the university deadline. This is the most beneficial advantage of the survival analysis approach that, in turn, leads to substantial gains in university costs.
In particular, in Figure 11 we report the results of Table 7. As we can observe, the Cox model for the faculty of Biology, CR(B), has a better performance in the first deciles (1-5). Considering the first 7 deciles the model captured the 81.67% of events for the faculty of psychology, CR(P), and the 87.09% of events for biology. The cumulative captured response rate is very important information for the university to plan the tutoring activity and to understand the weak points for specific exams.

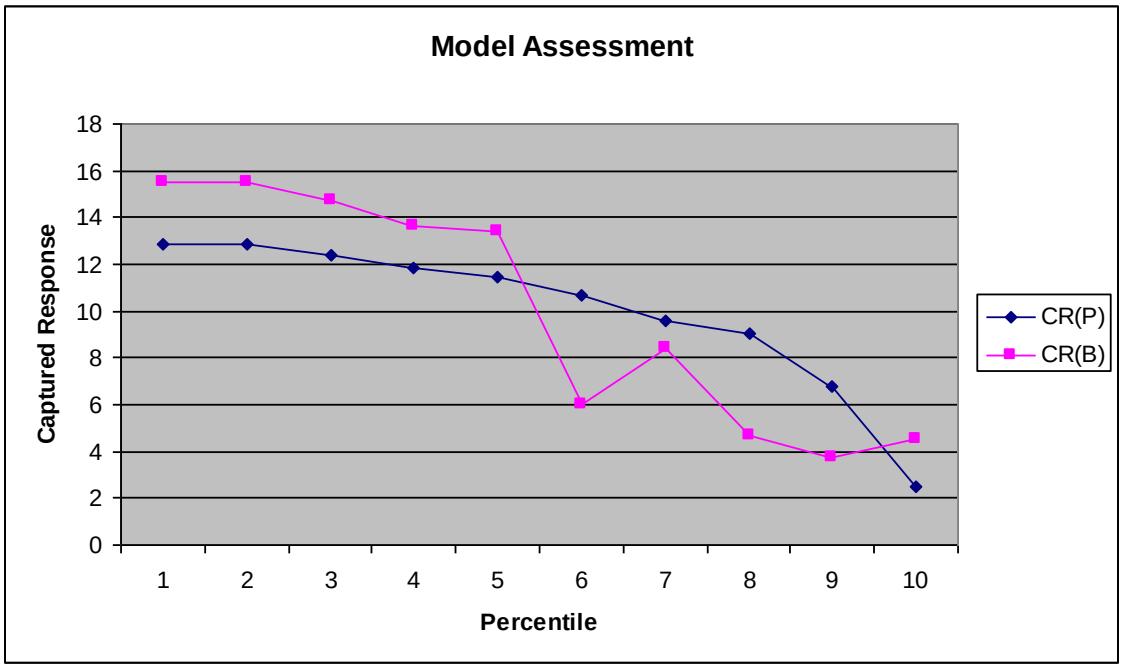

Figure 11: Cox model comparison

Log likelihood comparison can be formally embedded into an overall statistical test, such as Score, Wald or the likelihood ratio test. These tests compare the null assumption of no covariate effect against the alternative that at least one is different from zero.
From Table 8 and Table 9 we have further evidence on the effectiveness of fitting the chosen model, using either test. We underline that the final model contains 5 explanatory variables.
Considering the event of interest, based on the R square index, the final Cox model implemented explains 37.9% of the total variability for the Faculty of Psychology and 44,9% for the Faculty of Biology.

Table 8: Model diagnostics – Psychology

R-square 0.379		
Likelihood ratio test	402	$p < 0.0001$
Wald test	406	$p < 0.0001$
Score (logrank) test	529	$p < 0.0001$

Table 9: Model diagnostics - Biology

R-square 0.449		
Likelihood ratio test	618	$p < 0.0001$
Wald test	284	$p < 0.0001$
Score (logrank) test	454	$p < 0.0001$

5 Discussion and Conclusions

This paper address's the problem of student's evaluation. The objective reached with this research is a statistical methodology useful to study the phenomenon of churn and consequently to reduce the number of students who leave their university career without taking any degree.
More precisely, the methodological proposal considers, for the problem at hand, a static component (characteristics of the students), a dynamic component (trend and contacts of the students with the university), a seasonal part (period of exams) and external factors.
As pointed out in the paper, the use of the new models, which is based on survival analysis, is able to reduce the presence of fragmentary information, depending on the measurement time, and to improve predictive power.
The model proposed is able to define profiles of students with a high churn risk evaluating the dependence of risks on the basis of a relevant set of features collected. In order to reduce churn, the structures of the University in Pavia devoted to the actions of tutoring can choose specific groups of students, according to risk criteria and therefore of urgency.
The current research could be improved considering other faculties of the same athenaeum, or other universities in order to have terms of comparison. Furthermore, the methodology described in this paper is able also to derive a measure of churn risk for doctoral students.

References

[1] Anderson, P.K., Borgan, O., Gill, R.D., Keiding, N., 1993. Statistical Models based on Counting Processes. Springer, New York.
[2] Carnell, L., J., (2008), "The Effect of a Student-Designed Data Collection Project on Attitudes Toward Statistics," Journal of Statistics Education.
[3] Cox, D. R. and Oakes, D. (1984) Analysis of Survival Data. London: Chapman & Hall.
[4] Cox, D. R. (1972) Regression models and life-tables (with discussion). Journal of the Royal Statistical Society B 34, 187–220.
[5] Dutton, J., and Dutton, M. (2005), "Characteristics and Performance of Students in an Online Section of Business Statistics," Journal of Statistics Education.

[6] Dutton, J., Dutton, M. and Perry, J. (2001), "Do Online Students Perform as Well as Lecture Students?" Journal of Engineering Education, 90, 131-136.
[7] Fleming, T.R., Harrington, D.P., 1991. Counting Processes and Survival Analysis. Wiley, New York.
[8] Giudici, P. (2003) Applied data mining, Wiley:NY.
[9] Hurvich, C. M., and Tsai, C.-L. (1989), "Regression and Time Series Model Selection in Small Samples," Biometrika, 76, 297-307.
[10] Kaplan, E.L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53, 457-481.
[11] Simonoff, J.S., (1997), "The "Unusual Episode" and a Second Statistics Course," Journal of Statistics Education.
[12] Singer, J.D. and Willett, J.B. (2003) "Applied Longitudinal data analysis", Oxford University Press.
[13] Spooner, F., Jordan, L., Algozzine, B. and Spooner, M. (1999), "Student Ratings of Instruction in Distance Learning and On-Campus Classes," Journal of Educational Research, 92, 132-140.
[14] Wallace, D. R. and Mutooni, P. (1997), "A Comparative Evaluation of World Wide Web-Based and Classroom Teaching," Journal of Engineering Education, 86, 211 – 219.